様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ９ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2012～2015
課題番号：２４５００３６１

研究課題名（和文）タンパク質部分構造のモデル化による相互作用予測法

研究課題名（英文）Development of protein interaction prediction methods by modeling protein substructures

研究代表者
林田　守広（Hayashida, Morihiro）

京都大学・化学研究所・助教

研究者番号：４０４０２９２９

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：タンパク質の機能を知るために立体構造の理解が有用である．立体構造の類似度を測る指標として画像認識に使用されるSIFT局所特徴量記述子を利用した結果，これまで開発してきた画像圧縮による指標よりも優れていることを確認した．さらにタンパク質アミノ酸残基とRNA塩基との相互作用を予測するための条件付き確率場による新たなモデルを提案した．またタンパク質ヘテロ二量体，三量体を予測する，既存手法よりも極めて高精度な手法を開発した．

研究成果の概要（英文）：We developed a similarity measure between protein tertiary structures using SIFT local feature descriptor, which is often used in the field of image recognition. The proposed measure outperforms previously developed measures based on image compression. Furthermore, we proposed conditional random field models for predicting residue-base contacts in protein-RNA complexes. In addition, we developed methods for predicting protein heterodimers and heterotrimers in higher accuracy than existing methods.

研究分野：バイオインフォマティクス

キーワード：条件付き確率場　タンパク質RNA複合体　タンパク質複合体　SIFT特徴量

１．研究開始当初の背景

タンパク質は生体内で遺伝子を基に生成される，生命活動にとって重要な物質の一つである．個々の遺伝子の DNA 塩基配列に従って多種多様なタンパク質が存在し，それぞれある一定の立体構造をとることによって固有の機能を示す．特にタンパク質内部において固有の機能あるいは構造をもつ部分はドメインと呼ばれ，一つのタンパク質中に複数の異なるドメインを含むタンパク質が存在する一方，異なる複数のタンパク質に，同じ機能または構造をもつ特定のドメインが共通して含まれるものも存在する．

ドメインの定義は様々存在し，多数のアミノ酸配列からのプロファイル隠れマルコフモデルに基づく Pfam データベースや，二次構造を利用した SCOP データベース等がり，多くは専門家の手によって蓄積されてきた．

２．研究の目的

本研究の目的は，これまで探求してきたタンパク質立体構造間の高速な類似度指標の開発およびタンパク質相互作用予測法の開発をさらに推進し，タンパク質の詳細な内部構造を把握することでタンパク質の機能推定に役立てることである．

３．研究の方法

(1) タンパク質立体構造間の類似度指標を改良する．立体構造をアミノ酸残基の Cα原子間の距離行列として表現し，行列を画像とみなす．これまでの研究において，画像圧縮を利用した類似度指標として MSPC を開発した．しかしながら画像圧縮を利用した類似度計算では，タンパク質内部の類似する部分構造を直接抽出することは困難である．そこで高速かつ高性能な画像認識技術である SIFT (Scale Invariant Feature Transform) 局所特徴量記述子および，SIFT を近似高速化した SURF (Speeded Up Robust Features) 記述子を応用する．SIFT および SURF 記述子を用いることで，Cα原子間の距離行列における特徴点を抽出できる．立体構造間の類似度指標は，両方の距離行列において抽出した特徴点における SIFT または SURF 局所特徴量ベクトル間の距離に基づいて定義する．両方の距離行列において距離が最小となる特徴量ベクトルを見つけ，これらの距離の平均を類似度指標と定義する．

(2) タンパク質間相互作用の強度はタンパク質の機能性に係わっていることが知られているが実験による計測は困難である．転写因子複合体を構成するタンパク質の相互作用強度が弱い場合，標的遺伝子の転写は細胞内環境に強く依存すると考えられる．本研究では，相互作用強度を予測するために，タンパク質アミノ酸配列とドメイン領域からの新たな特徴量 SPD を提案する．SPD は，タンパク質アミノ酸配列のうち，ドメインとして認識されている領域に限定した spectrum kernel として定義される．予測問題は特徴量ベクトルと数値の関係を見つけるものであるので，教師付きの機械学習による回帰手法である SVR (Support Vector Regression)，RVM (Relevance Vector Machine)，およびオンライン回帰手法の ARCOR (Adaptive Regularization of weights for regression with COvariance Reset)，PA (Passive-Aggressive) アルゴリズムの回帰版を利用する．

(3) タンパク質と RNA の間の相互作用は，RNA のスプライシングや転写後調節，タンパク質翻訳等に係わっている．タンパク質 RNA 複合体の立体構造の解析が行われ，タンパク質がどのように RNA 塩基配列の特定部分を認識し結合しているのか研究がなされてきた．本研究では，タンパク質アミノ酸残基と RNA 塩基との相互作用を予測する手法を開発する．アミノ酸残基と塩基との関係を得るために，相互作用の進化的関係を利用する．相互作用する残基と塩基の一方が突然変異によって変化した場合，相互作用を維持するために，これに伴って塩基も変化する傾向があると考えられる．それ故多数のアミノ酸配列と塩基配列について，同一生物種における共起の度合いを相互情報量に基づいて計算する．この進化的な相関を入力として，隣接する残基および塩基の依存関係を条件付き確率場によってモデル化し，擬似対数尤度を最大化することによってモデルのパラメータを学習する．予測時には相互情報量と確率モデルから残基と塩基の相互作用確率を計算する．

(4) 多くのタンパク質は複数のタンパク質が結合し複合体を形成することによって初めて機能を示すと考えられている．酵母菌におけるタンパク質複合体を構成するタンパク質の数のうち，二つのタンパク質からなるタンパク質二量体は複合体全体の最も大きな割合を占める．これまで多数のタンパク質からなる複合体については，タンパク質相互作用ネットワークから複合体を予測する手法が開発されてきたが，多くは複合体内部の辺密度に依存するため二量体を予測するには不適切であった．本研究では，タンパク質二量体および三量体を予測するための，タンパク質相互作用ネットワークからの新たな特徴量とカーネル関数の開発を行う．

(5) タンパク質立体構造は距離の近い残基間に辺を張ったグラフとして表現することができる．グラフを効率よく圧縮する手法を開発することで類似部分構造の抽出に役立てる．一般のグラフの圧縮手法開発のためにまず，複数の木構造の圧縮を研究する．これまでに単一の木構造を構成する最小の文法を見つける手法を開発してきた．この手法を拡張し，複数の木構造を圧縮し，類似部分構造を見つける手法を開発する．

４．研究成果

(1) タンパク質の機能部位を立体構造の進化的に保存された部分構造から同定するために，高速な立体構造間の類似度指標を開発した．画像認識に用いられる SIFT および SURF 局所特徴量記述子に基づく類似度指標を新たに提案した．ベンチマークデータとして公開されているいくつかの構造分類されたタンパク質を用いて，二つのタンパク質が同じグループに分類されるか否かの二値分類として，AUC (Area Under receiver operating characteristic Curve) を算出した．その結果，SURF 記述子に基づく類似度指標は，これまで開発してきた MSPC と ACD あるいは NCD を組み合わせた類似度指標および既存の手法よりも，高い AUC を達成した．今後はさらに本研究による成果を改良し，距離行列に特化した指標の開発および類似部分構造の抽出を行う．またニューラルネットワーク等を用いた他の画像認識技術の応用も検討する．
(2) タンパク質間相互作用の強度を予測する新たな手法を開発した．タンパク質アミノ酸配列のドメイン領域に限定した spectrum kernel による特徴量 SPD を提案した．ベンチマークデータを用いた交差検定による性能評価を行った結果，既存手法の LPNM，ASNM，APM などよりも，SPD と SVR あるいは RVM を組み合わせた手法の方が平均予測誤差が小さく，さらに SPD とオンライン回帰手法 ARCOR，PA を組み合わせた手法の方がより平均予測誤差を小さくできることを示した．本研究により，提案する特徴量 SPD が相互作用強度予測に有用であることと教師付き回帰手法の有効性が示された．
(3) タンパク質アミノ酸残基と RNA 塩基の間の相互作用を予測する新たな確率モデルによる予測手法を開発した．特定の RNA に対するタンパク質アミノ酸配列の相互作用部位を予測する手法は存在したが，残基と塩基の相互作用を同時に予測する手法は前例がない．提案手法の有効性を検証するために，いくつかの立体構造が判明しているタンパク質 RNA 複合体を用いて，原子間の距離が 3Å 以内であるとき残基と塩基の間に相互作用があるとして交差検定を行った．条件付き確率場の入力として補正された相互情報量と残基，塩基のラベルを用いたモデルが最も高い AUC を示した．また 20 種のアミノ酸を 15 のグループへ分類したモデルが良い予測精度となった．さらに L1 ノルム正則化を行うことで予測精度の向上に成功した．
(4) タンパク質二量体および三量体を予測するための，信頼度付きタンパク質相互作用ネットワークからの新たな特徴量とタンパク質ドメイン構成に基づくカーネル関数を開発した．相互作用予測とは逆に，複合体を予測するためには複合体を構成しないタンパク質とは相互作用しないことを予測する必要がある．タンパク質内部のドメイン構成が同じであれば相互作用の有無も同じであると考え，カーネル関数を同一のドメイン構成であるときに 1，それ以外は 0 と定義した．提案手法を評価するために，タンパク質相互作用の信頼度として WI-PHI データ，複合体データとして CYC2008 を用い交差検定を行った．二量体予測，三量体予測それぞれにおいて，平均の F 値は既存の MCL，MCODE，RRW，NWE の手法よりも提案手法の方が極めて高い値を示した．三量体予測に関しては，一度パラメータを訓練データを使って学習した後に，予測対象の三つのタンパク質と重複するタンパク質の組について三量体かどうかの予測を行い，この三量体らしさを新たに特徴量として加える二段階での予測手法を提案し，二段階で予測することでさらに予測精度を向上させることができることを交差検定により確認した．これらの結果から提案するカーネル関数および特徴量がタンパク質二量体，三量体の予測に有効であることが示された．
(5) 複数の木構造を同時に生成する最小の木文法を見つけることで類似部分構造を抽出する手法を開発した．木文法は根付き木を対象とし，根での兄弟分割と内部頂点での親子分割を生成規則として持つ．複数の与えられた木構造のみを生成する最小の文法を見つける問題を整数計画問題へ定式化する方法を提案した．根付き木として表現できる糖鎖および RNA 二次構造を適当な方法で根付き木へ変換し，提案手法を適用した結果，類似木構造をいくつか見つけることができた他，階層的な構造も見つけることができた．

５．主な発表論文等

〔雑誌論文〕（計 9 件）

① Jira Jindalertudomdee, Morihiro Hayashida, Yang Zhao, Tatsuya Akutsu, Enumeration method for tree-like chemical compounds with benzene rings and naphthalene rings by breadth-first search order, BMC Bioinformatics, 査読有, 17 巻, 2016, 113
http://dx.doi.org/10.1186/s12859-016-0962-4

② Morihiro Hayashida, Jira Jindalertudomdee, Yang Zhao, Tatsuya Akutsu, Parallelization of enumerating tree-like chemical compounds by breadth-first search order, BMC Medical Genomics, 査読有, 8 巻, 2015, S15
http://dx.doi.org/10.1186/1755-8794-8-S2-S15

③ Yang Zhao, Morihiro Hayashida, Yue Cao, Jaewook Hwang, Tatsuya Akutsu, Grammar-based compression approach to extraction of common rules among multiple trees of glycans and RNAs, BMC Bioinformatics, 査読有, 16 巻, 2015, 128
http://dx.doi.org/10.1186/s12859-015-05

58-4
http://dx.doi.org/10.1186/1752-0509-7-S2-S15
④ Peiying Ruan, Morihiro Hayashida, Osamu Maruyama, Tatsuya Akutsu, Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels. BMC Bioinformatics, 査読有, 15 巻, 2014, S6
http://dx.doi.org/10.1186/1471-2105-15-S2-S6
⑤ Mayumi Kamada, Yusuke Sakuma, Morihiro Hayashida, Tatsuya Akutsu, Prediction of protein-protein interaction strength using domain features with supervised regression, The Scientific World Journal, 査読有, 2014 巻, 2014, 240673
http://dx.doi.org/10.1155/2014/240673
⑥ Morihiro Hayashida, Peiying Ruan, Tatsuya Akutsu, Proteome compression via protein domain compositions, Methods, 査読有, 67 巻, 2014, 380-385
http://dx.doi.org/10.1016/j.ymeth.2014.01.012
⑦ Morihiro Hayashida, Mayumi Kamada, Jiangning Song, Tatsuya Akutsu, Prediction of protein-RNA residue-base contacts using two-dimensional conditional random field with the lasso, BMC Systems Biology, 査読有, 7 巻, 2013, S15
⑧ Peiying Ruan, Morihiro Hayashida, Osamu Maruyama, Tatsuya Akutsu, Prediction of heterodimeric protein complexes from weighted protein-protein interaction networks using novel features and kernel functions, PLoS ONE, 査読有, 8 巻, 2013, e65265
http://dx.doi.org/10.1371/journal.pone.0065265
⑨ Yeuntyng Lai, Morihiro Hayashida, Tatsuya Akutsu, Survival analysis by penalized regression and matrix factorization, The Scientific World Journal, 査読有, 2013 巻, 2013, 632030
http://dx.doi.org/10.1155/2013/632030

〔学会発表〕(計 9 件)
① Morihiro Hayashida, Hitoshi Koyano, Integer linear programming approach to median and center strings for a probability distribution on a set of strings, The 7th International Conference on Bioinformatics Models, Methods and Algorithms, 2016/2/22, Rome, Italy
② Morihiro Hayashida, Mayumi Kamada, Hitoshi Koyano, Online learning approach to prediction of protein-protein interaction strengths, The 9th International Conference on Systems Biology, 2015/8/22, Luoyang, China
③ Morihiro Hayashida, Hitoshi Koyano, Tatsuya Akutsu, Measuring the similarity of protein structures using image local feature descriptors SIFT and SURF, The 8th International Conference on Systems Biology, 2014/10/26, Qingdao, China
④ Morihiro Hayashida, Jira Jindalertudomdee, Yang Zhao, Tatsuya Akutsu, Parallelization of enumerating tree-like chemical compounds by breadth-first search order, The 8th International Conference on Systems Biology, 2014/10/25, Qingdao, China
⑤ Peiying Ruan, Morihiro Hayashida, Tatsuya Akutsu, Study on weight of protein-protein interaction network for prediction of heterodimers, 電子情報通信学会総合大会, 2014/3/18, 新潟市
⑥ Peiying Ruan, Morihiro Hayashida, Osamu Maruyama, Tatsuya Akutsu, Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels, The 12th Asia Pacific Bioinformatics Conference, 2014/1/17, Shanghai, China
⑦ Morihiro Hayashida, Peiying Ruan, Tatsuya Akutsu, Proteome compression via protein domain compositions, 2013 IEEE Conference on Systems Biology, 2013/8/24, Huangshan, China
⑧ Yusuke Sakuma, Mayumi Kamada, Morihiro Hayashida, Tatsuya Akutsu, Inferring strengths of protein-protein interactions using support vector regression, The 2013 International Conference on Parallel and Distributed Processing Techniques and Applications, 2013/7/22, Las Vegas, USA
⑨ Morihiro Hayashida, Mayumi Kamada, Jiangning Song, Tatsuya Akutsu, Predicting protein-RNA residue-base contacts using two-dimensional conditional random field, 2012 IEEE Conference on Systems Biology, 2012/8/19, Xi'an, China

6．研究組織
(1)研究代表者
林田 守広（HAYASHIDA, Morihiro）
京都大学・化学研究所・助教
研究者番号：40402929